Genetic Networks of Complex Disorders: from a Novel Search Engine for PubMed Article Database

Genetic Networks of Complex Disorders: from a Novel Search Engine for PubMed Article Database

Jae-Yoon Jung, PhD[1], Dennis Paul Wall, PhD[1,2]
[1]Harvard Medical School, Boston, MA; [2]Beth Israel Deaconess Medical Center, Boston, MA

Abstract

Finding genetic risk factors of complex disorders may involve reviewing hundreds of genes or thousands of research articles iteratively, but few tools have been available to facilitate this procedure. In this work, we built a novel publication search engine that can identify target-disorder specific, genetics-oriented research articles and extract the genes with significant results. Preliminary test results showed that the output of this engine has better coverage in terms of genes or publications, than other existing applications. We consider it as an essential tool for understanding genetic networks of complex disorders.

Introduction

With advance of genotyping / sequencing technologies, a large volume of studies have reported genetic association with various disorders in the last decade, and the number is still increasing each year. As hundreds of genes may be involved in one complex disorder, a thorough literature review is a fundamental starting point to understand genetic risk factors of the target disorders. However, few applications have been available to help search and keep track of up-to-date, genetics-oriented articles. In this context, we aim to build a novel search engine that specifically focuses on identifying target disorders and associated human genes, from all original research articles in PubMed. The user-specified input for this engine is target disorder name(s), and the primary output is a list of candidate genes with significant results, supporting publications, and the main statements in such publications.

Methods

Given a user input of a target disorder, first we built an expanded PubMed search query with matching MeSH terms, disorder aliases, and publication type filters in order to retrieve disorder-specific research articles. Next, we selected genetics-oriented publications from them by searching for keywords in the title/abstract or MeSH, extracted from a training set of genetics research articles. Then we applied a rule-based text-mining algorithm to analyze title/abstract or MeSH terms, in order to identify 1) human gene symbols; 2) negation and structures in the title and abstract; and 3) characteristics of the study (e.g., linkage analysis, gene expression, genome-wide association, copy number variations, etc.). Finally, we identified the main candidate gene(s) per each publication using structural information obtained in the previous step. In addition, we assessed the collective significance of each candidate gene based on the number / importance of the related publications and the type of study.

Results

We tested our engine with five complex mental disorders (autism spectrum disorder, schizophrenia, bipolar disorder, major depression, and obsessive-compulsive disorder) and compared the results with those of existing databases that provide a list of candidate genes / publications per given disorder. In the case of autism spectrum disorder (ASD), for example, we found 12,900 original research articles (excluding news, comments, etc.), and 5,155 of them turned out to be genetics-related, including 959 reviews. About half of them (2,542) include names or symbols mapped to human genes, and we found 784 articles (excluding reviews) reporting 576 genes with significant test results in either the title or result/conclusion section of the abstract. Compared with results from SFARI (manually curated) and HuGE Navigator Phenopedia (algorithm based), our output data cover 212 more publications that include key reports of negative associations, and also include 14 significant candidate genes that are not present in these research sites.

Discussion

In this work, we implemented a novel publication search engine to find and summarize target articles, specifically focused on links between disorders and genotypes. Preliminary tests and comparisons with external databases of similar functionality showed that it can extensively search articles and correctly identify the main findings. We consider this engine as an essential tool for understanding and visualizing the genetic networks of complex disorders.